

- DOCUMENT RESUME

ED 412 239

TM 027 478

AUTHOR Butler, Olivia D.; Hanson, Bradley A.
TITLE Examination of Presmoothing and Postsmoothing Methods in
Equating a Direct Writing Assessment.
PUB DATE 1997-00-00
NOTE 15p.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Equated Scores; *Sample Size; Test Results; *Writing Tests
IDENTIFIERS *Equipercentile Equating; Linear Equating Method; *Smoothing
Methods

ABSTRACT

The effectiveness of smoothing in reducing random errors in equipercentile equating of a short writing assessment with two raters, two prompts, with scores ranging from zero to five was examined. Thirteen methods were examined: no equating, three presmoothing, three postsmoothing, three combination presmoothing and postsmoothing, mean equating, linear equating, and unsmoothed equipercentile. The data for the study resulted from simulations of a writing assessment with one and two raters used for a large testing program. Mean equating appears to have less error with small samples than the other methods. A combination of presmoothing and postsmoothing appears to have less error using a small sample with two raters. For the larger sample size, presmoothing with degree three appears to have less error than the other methods. Equating can be problematic with performance assessments that have small score ranged, however, it can be done and reduces error relative to no equating. (Contains 2 tables, 4 figures, and 11 references.) (Author/SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Examination of Presmoothing and Postsmoothing Methods in
Equating a Direct Writing Assessment

Olivia D. Butler
Wayne State University
Michigan Employment Security Agency

Bradley A. Hanson
American College Testing

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

Olivia Butler

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to
improve reproduction quality.

- Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

Abstract

The effectiveness of smoothing in reducing random errors in equipercentile equating of a short writing assessment with two raters, two prompts, with scores ranging from zero to five was examined. Thirteen methods were examined: no equating, 3 presmoothing, 3 postsmoothing, 3 combination presmoothing and postsmoothing, mean equating, linear equating, and unsmoothed equipercentile. The data for the study resulted from simulations of a writing assessment with one and two raters used for a large testing program. Mean equating appears to have less error with small samples than the other methods. A combination of presmoothing and postsmoothing appears to have less error using a small sample with two raters. For the larger sample size, presmoothing with degree 3 appears to have less error than the other methods. Equating can be problematic with performance assessments that have small score ranges, however it can be done and reduces error relative to no equating.

Acknowledgments: The authors wish to thank Deborah J. Harris and David M. Swarthout for their assistance.

Test equating is the process of removing an advantage obtained by those examinees who have been administered an easier form of a test. This is accomplished by creating equivalent scores across forms of the test. This concurs with Angoff's (1971, p. 563) equipercentile definition: "Two scores, one on form X and the other on form Y (where X and Y measure the same function with the same degree of reliability) may be considered equivalent if their corresponding percentile ranks in any given group are equal." After equating, either test can be used with confidence that the scores on the test are comparable.

Today, there is still concern by educators and researchers about the complexity of equating performance assessments in standardized testing situations, especially with writing assessments. This may be due to performance assessments usually having fewer items and score points as compared with traditional multiple choice exams. Several studies in the literature have focused on making scores comparable on direct writing assessments (writing sample data) (Harris & Welch, 1993; Phillips, 1985). However, relatively few studies examine or compare smoothing methods (presmoothing and postsmoothing), to increase the precision of equipercentile equating with performance assessment data. The aim of this study was to examine a combination of smoothing methods, to determine if their use will increase the accuracy of equating with performance assessment data.

This paper focuses on the random groups equating design. The random groups design in equating typically involves randomly assigning participants to the new and old forms of a test to be administered with a spiraling process. This spiraling process ensures that alternating examinees receive different forms of a test, to obtain randomly equivalent groups. Phillips (1989) used a random groups design with linear, polytomous Rasch, and equipercentile methods of equating, where test forms were spiraled within classrooms, resulting in randomly equivalent groups being administered each form.

Fairbank (1987), compared presmoothing and postsmoothing methods in equipercentile equating to increase the precision of the equating using simulated and operational data from an aptitude test. Hanson, Zeng, and Colton (1994), compared presmoothing and postsmoothing methods of equipercentile equating. The current study compared unsmoothed equipercentile equating, mean equating, linear equating, presmoothed and postsmoothed equipercentile equating, and combined presmoothed and postsmoothed methods using performance assessment data. Very little research has been conducted in using the combined smoothing methods to determine if equating precision is increased with performance assessment instruments. Usually either presmoothing or postsmoothing is done, not both.

Smoothing

This paper explores the performance of mean equating, linear equating, unsmoothed equipercentile, and smoothing equipercentile equating. Mean equating usually involves changes in the scale means. It may be considered an alternative to linear equating when using small samples

or when the standard deviations of two test forms are similar (Kolen, 1984). Linear equating allows for scores and means from two forms of an instrument to differ or vary along the score scale if they correspond to the same score. Equipercentile equating involves percentile ranks within a given distribution. Forms of an instrument are considered equivalent if they correspond to the same percentile rank (Harris, Welch, & Wang, 1994).

Smoothing can be used in equipercentile equating. There are two general categories of smoothing: presmoothing and postsmoothing. These methods have been the focus of several research studies regarding the quality of analytic smoothing (Hanson et al., 1994; Fairbank, 1987; Kolen, 1984). Presmoothing techniques are applied to the frequency distributions (raw score) before the equating procedure. Postsmoothing techniques are applied to the resulting conversions after an unsmoothed equipercentile equating of the forms. The intent of both methods is to reduce the amount of sampling error associated with sample dependent fluctuations from the data.

Determining what degrees of smoothing to use is subjective in nature. The degrees for smoothing selected for this study are based on commonly used degrees found in the literature. The models selected are: log-linear model for presmoothing and cubic splines for postsmoothing. The log-linear model for presmoothing is discussed by Holland & Thayer (1987) and Haberman (1974). This yields estimates of the raw score probabilities based on the maximum likelihood estimate of the parameters of the model given. Cubic splines for postsmoothing has been described by Kolen (1984) and Kolen & Jarjoura (1987).

Method

Equating error was estimated through simulation using two forms of a 10th grade writing assessment (Form A and Form B) using a program developed by the second author. There were 1,658 examinees who were administered Form A and 2,061 examinees who were administered Form B. Each form consisted of two prompts, one in the narrative mode and one in the explanatory mode. Responses to each prompt were scored by two raters on a 0 to 5 scale (the scores were integers from 0 to 5). Thus, the raw data for each form is contained in a $6 \times 6 \times 6 \times 6$ table (score of rater 1 on prompt 1 by score of rater 2 on prompt 1 by score of rater 1 on prompt 2 by score of rater 2 on prompt 2). Each cell of the 4-way table gives the number of examinees with a particular combination of the four scores. For each form a polynomial log-linear model was fit to the 4-way table (Haberman, 1974). Third degree polynomials were used for each marginal variable, and all first order 2-way, 3-way, and 4-way interactions were included in the model. The model fit the data well. The fitted distributions produced by the model were taken as population distributions and used to define population equating functions.

In this paper equating was performed for two different scores. One score was the sum of the scores on the two prompts for rater 1 (scores ranged from 0 to 10). The second score was the sum of scores on the two prompts from both raters (scores ranged from 0 to 20). For each score simulation results were obtained for two sample sizes: 250 per form, and 1,000 per form. For each score and each sample size the following steps were used for each replication of the

simulation.

1. Samples of the appropriate size are drawn from the population score distributions for Form A and Form B.
2. Thirteen equating functions, equating scores on Form B to scores on Form A, were computed: identity equating (no equating), mean equating, linear equating, unsmoothed equipercentile equating, equipercentile equating with polynomial log-linear presmoothing using degrees 6, 4, and 3 (three equatings), equipercentile equating with cubic spline postsmoothing with smoothing parameters 0.25, 0.50, and 0.75 (three equatings), equipercentile equating with polynomial log-linear presmoothing using degree 4 combined with cubic spline postsmoothing using smoothing parameters 0.25, 0.50, and 0.75 (three equatings).

Steps 1 and 2 were repeated 100 times producing 100 equating functions for each of the 13 equating methods. For each equating function at each raw score point the average squared difference between the equated score and the population equated score (computed using the population distributions) over the 100 replications was used as a measure of error (this is the mean squared error). The mean squared error (MSE) can be decomposed into two components: bias squared (systematic error), and variance (random error).

Results

Average values of MSE, squared bias, and variance were computed over raw score points using the population distribution of Form B to compute the average. These average MSE's, squared biases, and variances are reported in Table 1 for the score using 1 rater and Table 2 for the score using both raters. The column labeled "SE" in Tables 1 and 2 gives the standard error of the estimated average MSE. The abbreviations of the equating method used to label the rows in Tables 1 and 2 are: nequate (no equating), mean (mean equating), linear (linear equating), unsm (unsmoothed equipercentile equating), psm 6 (presmoothing with degree 6), psm 4 (presmoothing with degree 4), psm 3 (presmoothing with degree 3), po .25 (postsmoothing at .25), po .50 (postsmoothing at .50), po .75 (postsmoothing at .75), p4p.25 (presmoothing with degree 4 and postsmoothing at .25), p4p.50 (presmoothing with degree 4 and postsmoothing at .50), and p4p.75 (presmoothing with degree 4 and postsmoothing at .75).

Based on the results obtained with one rater for $n=250$, mean equating had less error (0.038181) than the other 11 methods. However, looking at the same sample size for presmoothing at degrees 6, 4, and 3, presmoothing at 3 degrees has less error (0.046079) than degrees 6 (0.054079) and 4 (0.049830). For the postsmoothing methods, postsmoothing at .25 had less error although the differences among postsmoothing at .25, .50, and .75 are minimal. This is also true for the combination of presmoothing and postsmoothing methods. The difference in the MSE is minimal among the combination methods. However, the combination of presmoothing with degree 4 and postsmoothing at .25 resulted in less error. For sample $n=1,000$,

presmoothing with degree 3 appears to be the best method with MSE being 0.009990. It has more random than systematic error but the amount is small (0.009686). Figures 1 and 2 show the MSE at each raw score point for sample sizes of 250 and 1,000, respectively.

Table 2 shows the average results for the scores using two raters, with sample sizes 250 and 1,000. For the small sample of $n=250$ the combination of presmoothing 4 and postsmoothing .25 showed the least error overall 0.161252. For the sample of $n=1,000$, presmoothing with degree 3 had the least amount of error 0.037661. Figures 3 and 4 show the MSE at each raw score for sample sizes of 250 and 1,000, respectively. In samples of 1,000, with one and two raters presmoothing with degree 3 showed the least amount of error.

It does not appear that combining smoothing methods made a difference except for the sample size of 250 with one rater. The results indicate that it is better to do equating than no equating. Tables 1 and 2 show the average MSE associated with no equating, is larger than the MSE for any of the other equating methods. It is also noted that the error rates are much smaller using larger sample sizes with both one and two raters.

Conclusion

Equating procedures are all statistical techniques that are subject to random and systematic error. There is less random error associated with linear equating than with equipercentile equating. However, this study showed that smoothing using performance assessment data can be beneficial. Even with small sample sizes, few score points and one rater, equating worked better than no equating. Presmoothing worked best for the larger sample size with one and two raters. Both methods of smoothing show evidence that these methods can reduce random equating error more than they increase systematic equating error.

The results indicated that combining presmoothing and postsmoothing did not reduce equating error appreciably compared with using presmoothing or postsmoothing alone. The results further indicated that smoothing can reduce overall error in equipercentile equating. Either presmoothing or postsmoothing can be used, but no additional benefit is obtained by performing both presmoothing and postsmoothing.

Based on this study, a recommendation on the best equating method could not be determined because the results varied for the different conditions. For example, with a small sample size ($n=250$) and one rater, mean equating resulted in less error. The same sample size with two raters resulted in the combination of presmoothing 4 and postsmoothing .25 being the best. However, for the larger sample sizes with one and two raters, presmoothing with degree 3 had the smallest error. Of course, larger sample sizes are always preferred because the random error is smaller.

One of several limitations with this study is that it is unclear to what extent the results will generalize to other situations. Also, one should evaluate the appropriateness of the models used

for equating. Additional research on equating with performance assessments should be pursued. For example, how do the results for performance assessments compare with other tests or other performance assessment instruments used operationally? It is hoped that this research will motivate additional questions and research on equating performance assessments.

Table 1

SIMULATED SAMPLE - ONE RATER

n=250

<u>Method</u>	<u>Squared Bias</u>	<u>Variance</u>	<u>MSE</u>	<u>SE</u>
nequate	0.130002	0.000000	0.130002	0.000000
mean	0.014724	0.023456	0.038181	0.003177
linear	0.008103	0.040619	0.048722	0.003986
unsm	0.001336	0.054660	0.055996	0.003753
psm 6	0.001565	0.052515	0.054079	0.003738
psm 4	0.001491	0.048340	0.049830	0.003760
psm 3	0.001524	0.044554	0.046079	0.003759
po .25	0.004019	0.041268	0.045287	0.003877
po .50	0.005218	0.040525	0.045742	0.004022
po .75	0.005936	0.040118	0.046054	0.004081
p4p.25	0.004774	0.040024	0.044798	0.003897
p4p.50	0.005923	0.039873	0.045795	0.003984
p4p.75	0.006474	0.039899	0.046373	0.004019

n=1,000

nequate	0.130002	0.000000	0.130002	0.000000
mean	0.014192	0.005168	0.019360	0.000756
linear	0.006916	0.008707	0.015623	0.001032
unsm	0.000105	0.013259	0.013365	0.001010
psm 6	0.000258	0.012090	0.012348	0.001015
psm 4	0.000281	0.010775	0.011057	0.000993
psm 3	0.000304	0.009686	0.009990	0.000987
po .25	0.002130	0.009450	0.011581	0.000969
po .50	0.002874	0.009321	0.012195	0.000969
po .75	0.003435	0.009223	0.012658	0.000984
p4p.25	0.002871	0.008990	0.011862	0.000980
p4p.50	0.003689	0.008822	0.012511	0.000995
p4p.75	0.004139	0.008766	0.012905	0.001021

Table 2

SIMULATED SAMPLE - TWO RATERS

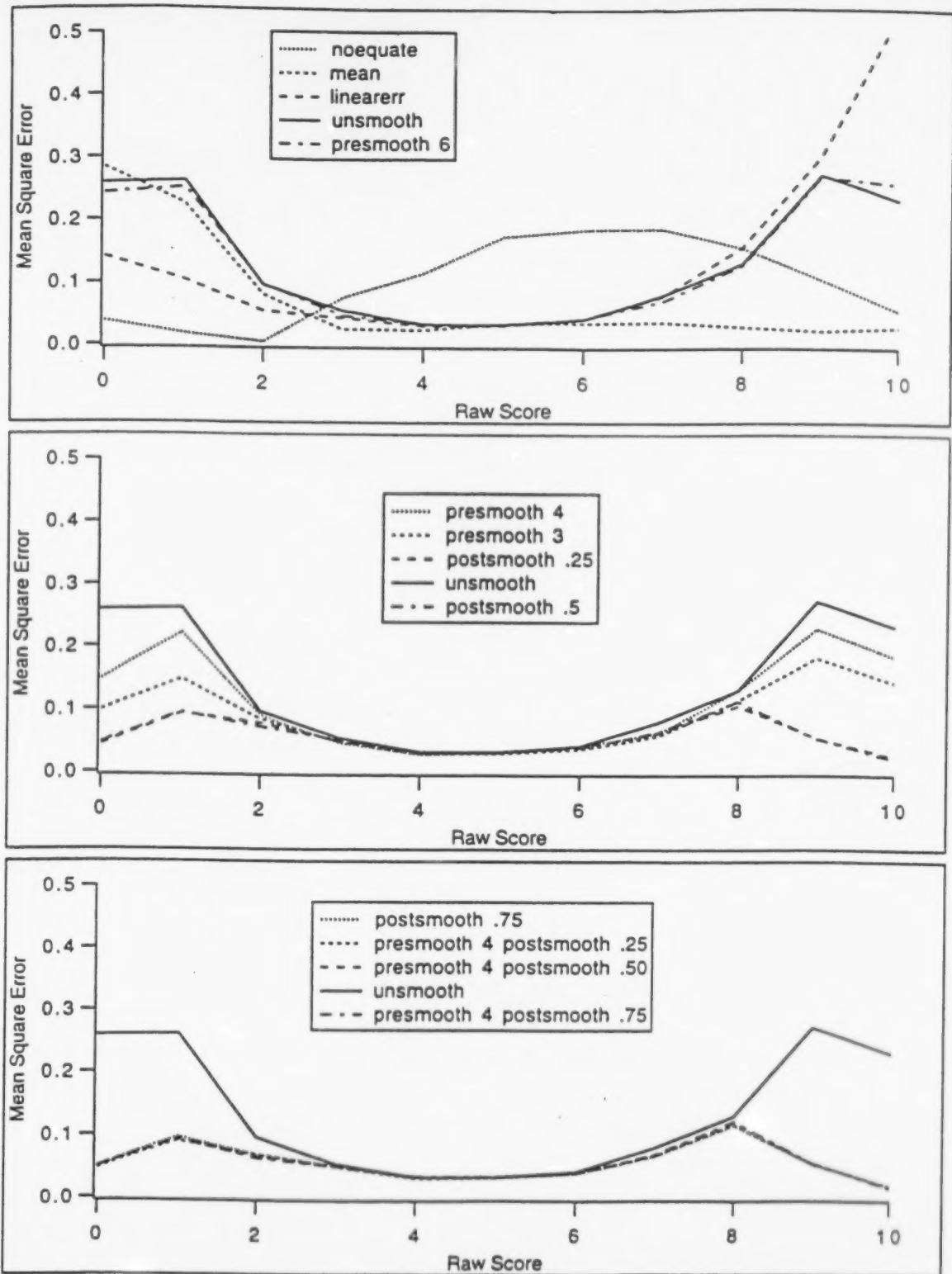
n=250

<u>Method</u>	<u>Squared Bias</u>	<u>Variance</u>	<u>MSE</u>	<u>SE</u>
nequate	0.539115	0.000000	0.539115	0.000000
mean	0.083405	0.078539	0.161944	0.011272
linear	0.022001	0.148437	0.170438	0.015508
unsm	0.007422	0.222576	0.229998	0.015530
psm 6	0.007345	0.206362	0.213707	0.015719
psm 4	0.006985	0.188852	0.195838	0.015561
psm 3	0.006476	0.170010	0.176486	0.015119
po .25	0.009152	0.162569	0.171720	0.015118
po .50	0.012836	0.151903	0.164738	0.015364
po .75	0.016309	0.147910	0.164219	0.015789
p4p.25	0.012678	0.148574	0.161252	0.015173
p4p.50	0.015901	0.146550	0.162451	0.015430
p4p.75	0.017486	0.145765	0.163250	0.015523

n=1,000

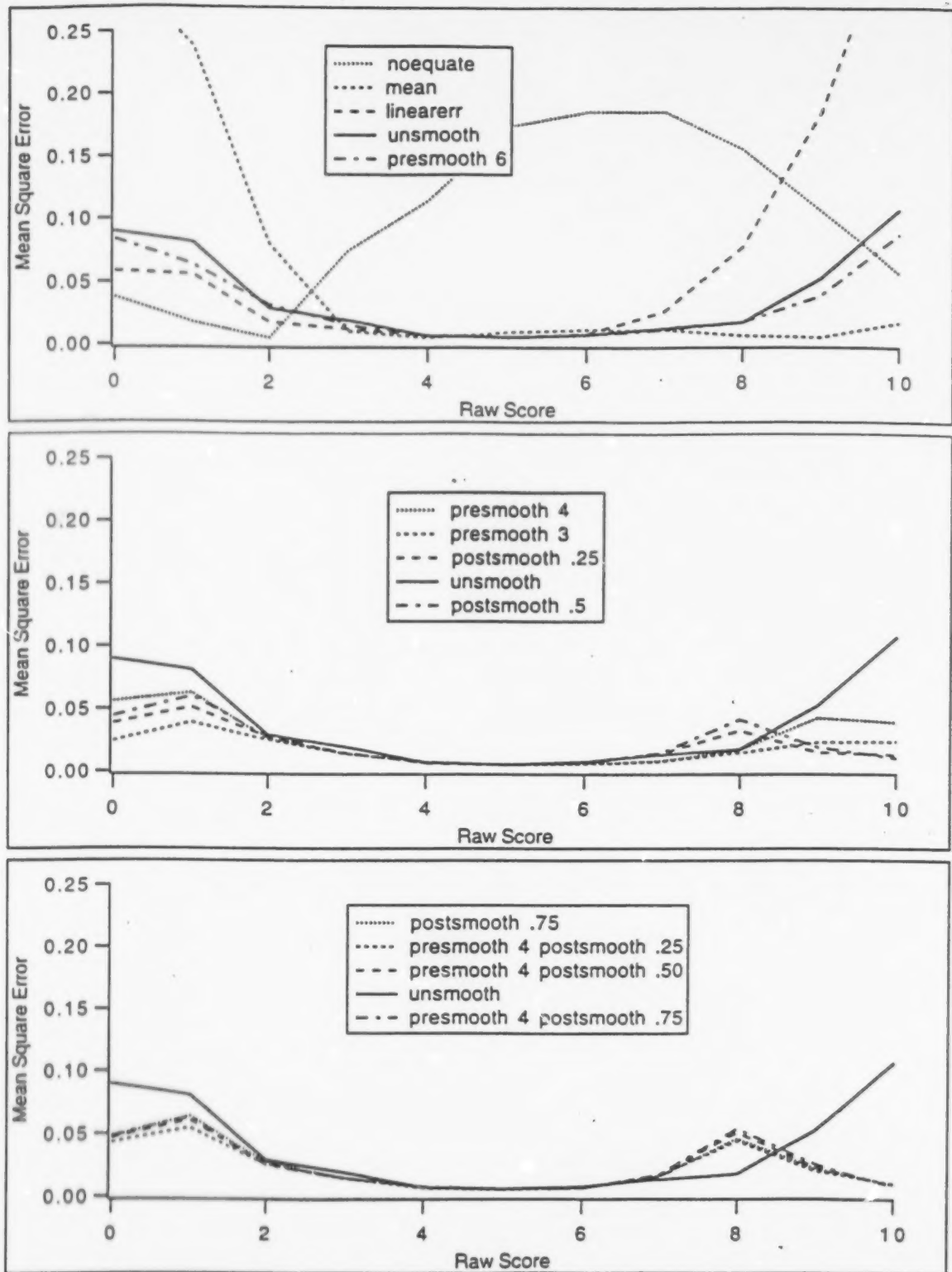
nequate	0.539115	0.000000	0.539115	0.000000
mean	0.079039	0.017935	0.096974	0.002770
linear	0.016854	0.030002	0.046856	0.003415
unsm	0.000611	0.051797	0.052408	0.003585
psm 6	0.001201	0.045838	0.047039	0.003619
psm 4	0.001326	0.040207	0.041533	0.003515
psm 3	0.001375	0.036285	0.037661	0.003494
po .25	0.002588	0.036678	0.039266	0.003583
po .50	0.005287	0.034313	0.039600	0.003583
po .75	0.007951	0.033474	0.041424	0.003638
p4p.25	0.006985	0.032571	0.039556	0.003497
p4p.50	0.009994	0.031628	0.041622	0.003560
p4p.75	0.011369	0.031274	0.042643	0.003574

Figure 1



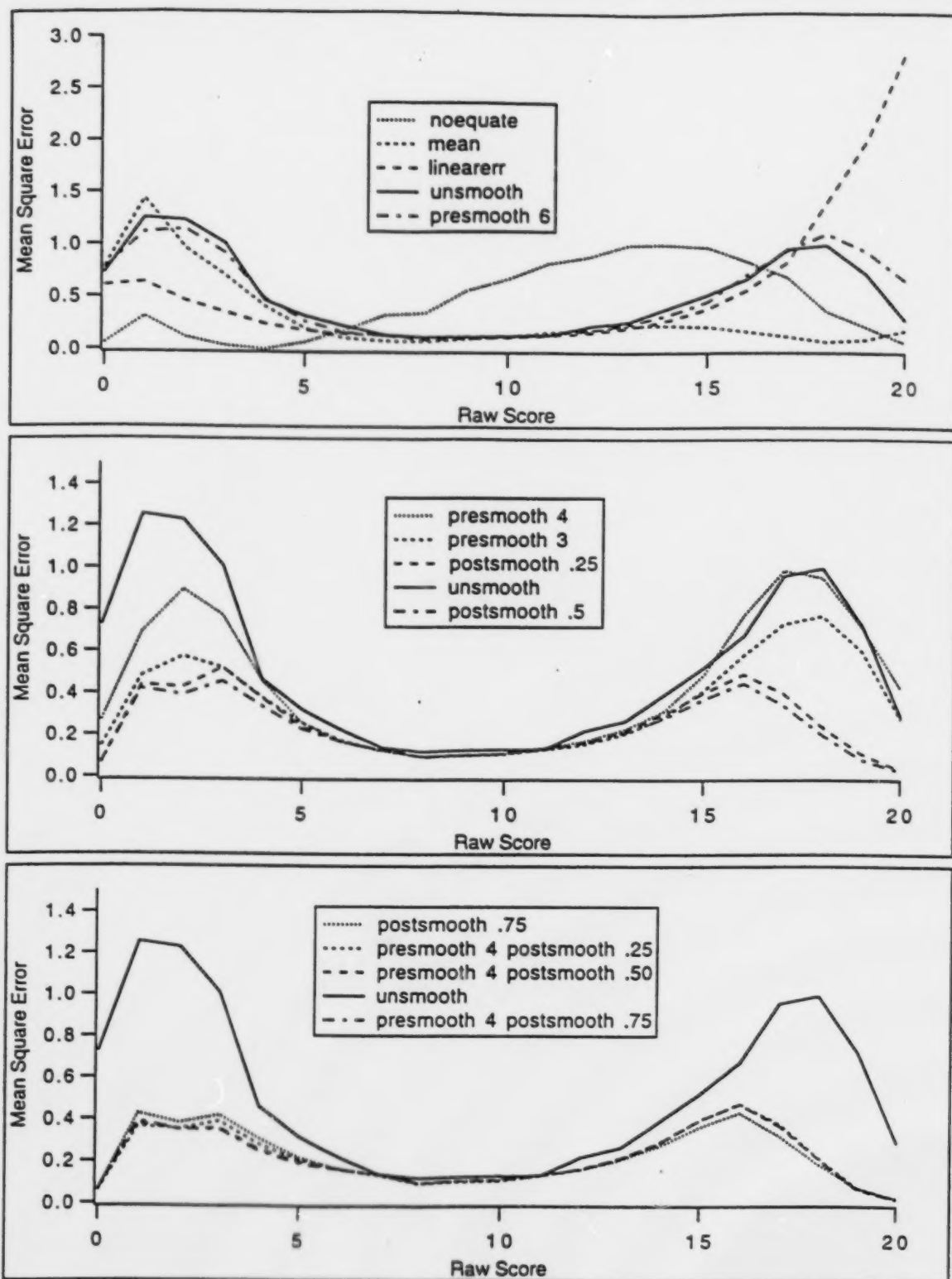
Equating Functions for Simulated PLAN Writing Test, N=250, 1 Rater

Figure 2



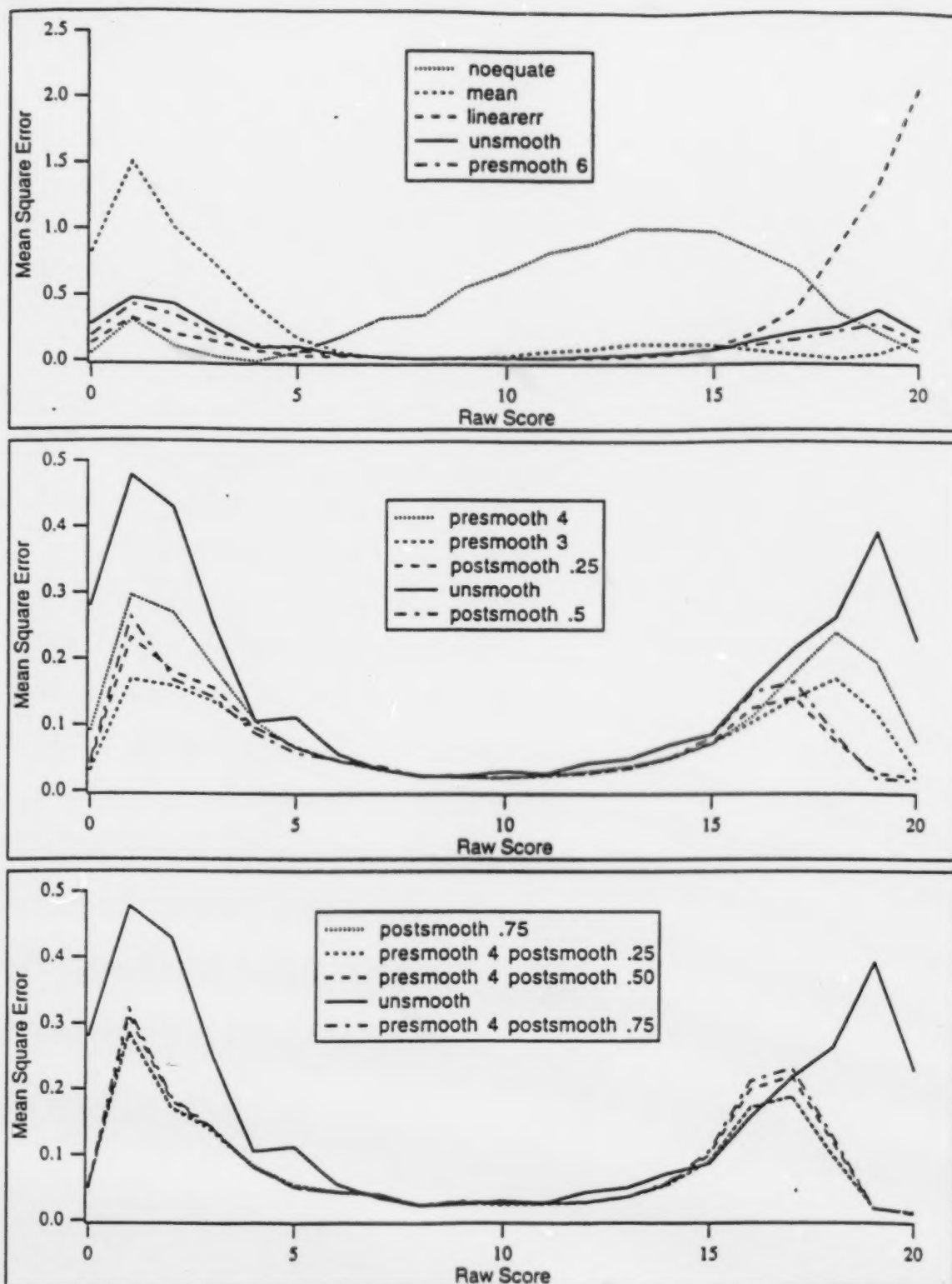
Equating Functions for Simulated PLAN Writing Test, N=1000, 1 Rater

Figure 3



Equating Functions for Simulated PLAN Writing Test, N=250, 2 Raters

Figure 4



Equating Functions for Simulated PLAN Writing Test, N=1000, 2 Raters

References

- ACT (1994). 10th grade writing assessment. Iowa City, IA: American College Testing.
- Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.
- Fairbank, B.A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement*, 11, 245-262.
- Haberman, S.J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics*, 30, 589-600.
- Hanson, B.A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (Research Report 94-4). Iowa City, IA: American College Testing.
- Harris, D.J., & Welch, C.J. (1993). *Equating writing samples*. Paper presented at the Annual Meeting of the National Council on Measurement in Education.
- Harris, D.J., Welch, C.J., & Wang, T. (1994). *Issues in equating writing assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education.
- Holland, P.W. & Thayer, D.T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Educational Testing Service Research Report 87-31). Princeton, NJ: Educational Testing Service.
- Kolen, M.J. (1984). Effectiveness of analytical smoothing in equipercentile equating. *Journal of Educational Statistics*, 9, 25-44.
- Kolen, M.J. & Brennan, R.L. (1995). *Test equating: methods and practices*. New York, NY: Springer-Verlag.
- Kolen, M.J., & Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometrika*, 52, 43-59.